# explicação

# 📋 Relatório Final - Engenharia do Conhecimento 2024/2025

## 📚 Passo 1: Análise Exploratória de Dados (EDA)

### 1.0 Importação de Bibliotecas

- Bibliotecas de manipulação de dados, visualização e machine learning foram importadas antecipadamente para uso posterior.

### 1.1 Carregamento de Dados

- Dados carregados do arquivo `custom_covid19.csv`.
- Visualização das 5 primeiras linhas para uma inspeção inicial.

### 1.2 Identificação dos Tipos de Variáveis

- Variáveis categóricas, numéricas e especiais (como `DATE_DIED`) foram identificadas.
- Nova coluna `DIED` criada (0 = vivo, 1 = falecido).

### 1.3 Análise de Valores Ausentes

- Valores 97, 98 e 99 tratados como ausentes.
- Gráfico de barras gerado para visualizar a distribuição de valores ausentes.

### 1.4 Detecção de Outliers

- Método IQR aplicado para detectar outliers em variáveis numéricas (principalmente `AGE`).
- Gráficos Boxplot gerados.

### 1.5 Distribuição das Variáveis

- Gráficos de distribuição e contagem gerados para variáveis importantes como `AGE`, `SEX`, `ICU`, etc.

### 1.6 Análise de Correlação

- Matriz de correlação gerada entre variáveis numéricas.
- Heatmap de correlação visualizado.

### 1.7 Reconhecimento de Variáveis Importantes

- Com base na análise estatística e conhecimento médico, foram selecionadas variáveis relevantes para:
  - O1 (Previsão de morte)
  - O2 (Previsão de idade)
  - O3 (Previsão da idade entre falecidos)

### 1.8 Comparação de Médias por Grupos

- Comparação dos valores médios de variáveis importantes entre vivos e falecidos.

### 1.9 Baseline - Árvore de Decisão

- Treinamento de uma árvore de decisão simples como baseline para a tarefa O1.

### 1.10 Tratamento de Valores Ausentes

- Preenchimento de valores ausentes:
  - Categóricos com moda.
  - Numéricos com mediana.

### 1.11 Salvamento do Dataset Limpo

- Correção de tipos de dados.
- Salvamento do dataset limpo em `custom_covid19_cleaned.csv`.

---

## 🎯 Passo 2: Tarefa O1 - Previsão de Morte (Classificação)

### 2.1 Preparação dos Dados

- Seleção das variáveis importantes para O1.
- Separação em conjunto de treino e validação.

### 2.2 Treinamento de Modelos

- Treinamento de:

- Árvore de Decisão (baseline)
- Regressão Logística (com normalização dos dados)

### 2.3 Avaliação dos Modelos

- Avaliação usando métricas como F1-score, MCC e matrizes de confusão.
- Regressão Logística apresentou melhor desempenho que Árvore de Decisão.

### 2.4 Comparação e Seleção do Melhor Modelo

- Regressão Logística escolhida como base para otimização.

### 2.5 Otimização da Regressão Logística

- Aplicação de GridSearchCV para otimizar o hiperparâmetro `C`.
- Melhor modelo salvo como `logistic_regression_best.pkl`.

### 2.6 Ajuste da Árvore de Decisão (Opcional)

- Teste de diferentes hiperparâmetros via GridSearch, mas sem superação da Regressão Logística.

---

## 🎯 Passo 3: Tarefa O2 - Previsão da Idade (Regressão)

### 3.1 Preparação dos Dados

- Seleção das variáveis relevantes para previsão de idade.
- Padronização dos dados.

### 3.2 Treinamento de Modelos

- Treinamento de:
  - Random Forest Regressor
  - Regressão Ridge
  - Regressão Linear
- SVR descartado devido ao tempo de treinamento excessivo.

### 3.3 Avaliação dos Modelos

- Avaliação usando MAE, RMSE, $R^2$ e correlação de Pearson.

- Random Forest apresentou melhor desempenho geral.

### 3.4 Visualização

- Gráfico de dispersão Real vs Predito para Random Forest.

### 3.5 Salvamento do Melhor Modelo

- Random Forest salvo como `random_forest_age_predictor.pkl`.

---

## 🎯 Passo 4: Tarefa O3 - Previsão da Idade entre Pacientes Falecidos

### 4.1 Preparação dos Dados

- Seleção apenas dos pacientes falecidos.
- Seleção das variáveis relevantes (important_features_O3).
- Padronização dos dados.

### 4.2 Treinamento do Modelo

- Random Forest treinado para prever idade entre falecidos.

### 4.3 Avaliação

- Avaliação dos erros e coeficientes.
- Resultado demonstrou que prever idade apenas com essas variáveis é extremamente difícil.

### 4.4 Salvamento do Modelo

- Modelo salvo como `random_forest_deceased_age_predictor.pkl`.

---

## 🧠 Passo 5: Resumo dos Atributos e Conclusões

### 5.1 Resumo das Principais Variáveis

- **O1 (Previsão de morte)**: idade, ICU, intubação, pneumonia, diabetes, etc.
- **O2 (Previsão da idade)**: diabetes, tabagismo, doenças crônicas.
- **O3 (Previsão da idade entre falecidos)**: características de gravidade e comorbidades.

## 5.2 Tabela de Importância dos Atributos

| Tarefa | Principais Variáveis | Observações |
|---|---|---|
| O1 (Previsão de morte) | idade, ICU, intubação, pneumonia, diabetes, etc. | Idade e gravidade da condição têm grande impacto na mortalidade. |
| O2 (Previsão da idade) | doenças crônicas, tabagismo | Associação moderada com a idade. |
| O3 (Previsão da idade entre falecidos) | condições graves | Difícil prever idade com base apenas em comorbidades. |

## 5.3 Observações Finais

- A tarefa O1 foi relativamente mais fácil e os modelos tiveram bom desempenho.
- A tarefa O2 teve desempenho razoável; a relação entre doenças e idade é fraca.
- A tarefa O3 mostrou ser extremamente difícil de resolver com os dados disponíveis.

---

**Fim do Relatório**